United Nations A/HRC/38/35

General Assembly

Distr.: General
6 April 2018

Original: English

Human Rights Council
Thirty-eighth session
18 June–6 July 2018
Agenda item 3
Promotion and protection of all human rights, civil, political, economic, social and cultural rights, including the right to development

Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression

Note by the Secretariat

The Secretariat has the honour to transmit to the Human Rights Council the report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, David Kaye, pursuant to Council resolution 34/18. In his report the Special Rapporteur addresses the regulation of user-generated online content. He recommends that States ensure an enabling environment for online freedom of expression and that companies apply human rights standards at all stages of their operations. Human rights law gives companies the tools to articulate their positions in ways that respect democratic norms and counter authoritarian demands. At a minimum, companies and States should pursue radically improved transparency, from rule-making to enforcement of the rules, to ensure user autonomy as individuals increasingly exercise fundamental rights online.

GE.18-05436(E)
1805436

Please recycle

Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression

Contents

		Page
I.	Introduction	3
II.	Legal framework	3
	A. State obligations	4
	B. Company responsibilities	5
III.	Key concerns with content regulation	6
	A. Government regulation	6
	B. Company moderation of content	8
IV.	Human rights principles for company content moderation	14
	A. Substantive standards for content moderation	15
	B. Processes for company moderation and related activities	16
V.	Recommendations	19

I. Introduction

1. Early in the digital age, John Perry Barlow declared that the Internet would usher in "a world where anyone, anywhere may express his or her beliefs, no matter how singular, without fear of being coerced into silence or conformity".[1] Although the Internet remains history's greatest tool for global access to information, such online evangelism is hard to find today. The public sees hate, abuse and disinformation in the content users generate. Governments see terrorist recruitment or discomfiting dissent and opposition. Civil society organizations see the outsourcing of public functions, like protection of freedom of expression, to unaccountable private actors. Despite taking steps to illuminate their rules and government interactions, the companies remain enigmatic regulators, establishing a kind of "platform law" in which clarity, consistency, accountability and remedy are elusive. The United Nations, regional organizations and treaty bodies have affirmed that offline rights apply equally online, but it is not always clear that the companies protect the rights of their users or that States give companies legal incentives to do so.

2. In the present report the Special Rapporteur proposes a framework for the moderation of user-generated online content that puts human rights at the very centre.[2] He seeks to answer basic questions: What responsibilities do companies have to ensure that their platforms do not interfere with rights guaranteed under international law? What standards should they apply to content moderation? Should States regulate commercial content moderation and, if so, how? The law expects transparency and accountability from States to mitigate threats to freedom of expression. Should we expect the same of private actors? What do the processes of protection and remedy look like in the digital age?

3. Previous reports have addressed some of these questions.[3] The present report focuses on the regulation of user-generated content, principally by States and social media companies but in a way that is applicable to all relevant actors in the information and communications technology (ICT) sector. The Special Rapporteur outlines the applicable human rights legal framework and describes company and State approaches to content regulation. He proposes standards and processes that companies should adopt to regulate content in accordance with human rights law.

4. Research into the companies' terms of service, transparency reporting and secondary sources provided the initial basis for the report. Calls for comments generated 21 submissions from States and 29 from non-State actors (including 1 company submission). The Special Rapporteur visited several companies in Silicon Valley and held conversations with others in an effort to understand their approaches to content moderation.[4] He benefited from civil society consultations held in Bangkok and Geneva in 2017 and 2018 and online discussions with experts in Latin America, the Middle East and North Africa and sub-Saharan Africa in 2018.[5]

II. Legal framework

5. The activities of companies in the ICT sector implicate rights to privacy, religious freedom and belief, opinion and expression, assembly and association, and public participation, among others. The present report focuses on freedom of expression while

[1] John Perry Barlow, A Declaration of the Independence of Cyberspace, 8 February 1996.

[2] "Moderation" describes the process by which Internet companies determine whether user-generated content meets the standards articulated in their terms of service and other rules.

[3] A/HRC/35/22 and A/HRC/32/38.

[4] The Special Rapporteur visited the headquarters of Facebook, Github, Google, Reddit and Twitter and held conversations with representatives of Yahoo/Oath, Line and Microsoft. He also visited the non-profit Wikimedia Foundation. He hopes to visit companies in Beijing, Moscow, Seoul and Tokyo in work related to the present report.

[5] The Special Rapporteur wishes to thank his legal adviser, Amos Toh, and students at the International Justice Clinic at the University of California, Irvine, School of Law.

acknowledging the interdependence of rights, such as the importance of privacy as a gateway to freedom of expression.[6] Article 19 of the International Covenant on Civil and Political Rights provides globally established rules, ratified by 170 States and echoing the Universal Declaration of Human Rights, guaranteeing "the right to hold opinions without interference" and "the right to seek, receive and impart information and ideas of all kinds, regardless of frontiers" and through any medium.[7]

A. State obligations

6. Human rights law imposes duties on States to ensure enabling environments for freedom of expression and to protect its exercise. The duty to ensure freedom of expression obligates States to promote, *inter alia*, media diversity and independence and access to information.[8] Additionally, international and regional bodies have urged States to promote universal Internet access.[9] States also have a duty to ensure that private entities do not interfere with the freedoms of opinion and expression.[10] The Guiding Principles on Business and Human Rights, adopted by the Human Rights Council in 2011, emphasize in principle 3 State duties to ensure environments that enable business respect for human rights.[11]

7. States may not restrict the right to hold opinions without interference. Per article 19 (3) of the Covenant, State limitations on freedom of expression must meet the following well-established conditions:

- *Legality*. Restrictions must be "provided by law". In particular, they must be adopted by regular legal processes and limit government discretion in a manner that distinguishes between lawful and unlawful expression with "sufficient precision". Secretly adopted restrictions fail this fundamental requirement.[12] The assurance of legality should generally involve the oversight of independent judicial authorities.[13]
- *Necessity and proportionality*. States must demonstrate that the restriction imposes the least burden on the exercise of the right and actually protects, or is likely to protect, the legitimate State interest at issue. States may not merely assert necessity but must demonstrate it, in the adoption of restrictive legislation and the restriction of specific expression.[14]
- *Legitimacy*. Any restriction, to be lawful, must protect only those interests enumerated in article 19 (3): the rights or reputations of others, national security or public order, or public health or morals. Restrictions designed to protect the rights of others, for instance, include "human rights as recognized in the Covenant and more generally in international human rights law".[15] Restrictions to protect rights to privacy, life, due process, association and participation in public affairs, to name a few, would be legitimate when demonstrated to meet the tests of legality and necessity. The Human Rights Committee cautions that restrictions to protect "public

---

[6] See A/HRC/29/32, paras. 16–18.

[7] See also African Charter on Human and Peoples' Rights, art. 9; American Convention on Human Rights, art. 13; Convention for the Protection of Human Rights and Fundamental Freedoms, art. 10. See also Centro de Estudios en Libertad de Expresión y Acceso a la Información submission.

[8] Joint Declaration on Freedom of Expression and "Fake News", Disinformation and Propaganda, 3 March 2017, sect. 3. See also Human Rights Committee, general comment No. 34 (2011) on the freedoms of opinion and expression, paras. 18 and 40; A/HRC/29/32, para. 61 and A/HRC/32/38, para. 86.

[9] See Human Rights Council, resolution 32/13, para. 12; Office of the Special Rapporteur for Freedom of Expression of the Inter-American Commission on Human Rights, *Standards for a Free, Open and Inclusive Internet* (2016), para. 18.

[10] See general comment No. 34, para. 7.

[11] A/HRC/17/31.

[12] Ibid. para. 25; A/HRC/29/32.

[13] Ibid.

[14] See general comment No. 34, para. 27.

[15] Ibid., para. 28.

morals" should not derive "exclusively from a single tradition", seeking to ensure that the restriction reflects principles of non-discrimination and the universality of rights.[16]

8. Restrictions pursuant to article 20 (2) of the Covenant — which requires States to prohibit "advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence" — must still satisfy the cumulative conditions of legality, necessity and legitimacy.[17]

B. Company responsibilities

9. Internet companies have become central platforms for discussion and debate, information access, commerce and human development.[18] They collect and retain the personal data of billions of individuals, including information about their habits, whereabouts and activities, and often claim civic roles. In 2004, Google promoted its ambition to do "good things for the world even if we forgo some short term gains".[19] Facebook's founder has proclaimed a desire to "develop the social infrastructure to give people the power to build a global community that works for all of us".[20] Twitter has promised policies that "improve — and do not detract from — a free and global conversation".[21] VKontakte, a Russian social media company, "unites people all over the world", while Tencent reflects the language of the Government of China when noting its aims to "help build a harmonious society and to become a good corporate citizen".[22]

10. Few companies apply human rights principles in their operations, and most that do see them as limited to how they respond to government threats and demands.[23] However, the Guiding Principles on Business and Human Rights establish "global standard[s] of expected conduct" that should apply throughout company operations and wherever they operate.[24] While the Guiding Principles are non-binding, the companies' overwhelming role in public life globally argues strongly for their adoption and implementation.

11. The Guiding Principles establish a framework according to which companies should, at a minimum:

(a) Avoid causing or contributing to adverse human rights impacts and seek to prevent or mitigate such impacts directly linked to their operations, products or services by their business relationships, even if they have not contributed to those impacts (principle 13);

(b) Make high-level policy commitments to respect the human rights of their users (principle 16);

(c) Conduct due diligence that identifies, addresses and accounts for actual and potential human rights impacts of their activities, including through regular risk and impact assessments, meaningful consultation with potentially affected groups and other stakeholders, and appropriate follow-up action that mitigates or prevents these impacts (principles 17–19);

(d) Engage in prevention and mitigation strategies that respect principles of internationally recognized human rights to the greatest extent possible when faced with conflicting local law requirements (principle 23);

---

[16] Ibid., para. 32.

[17] Ibid., para. 50. See also A/67/357.

[18] See, for example, Supreme Court of the United States, *Packingham v. North Carolina*, opinion of 19 June 2017; European Court of Human Rights, *Times Newspapers Ltd. (Nos. 1 and 2) v. The United Kingdom* (application Nos. 3002/03 and 23676/03), judgment of 10 March 2009, para. 27.

[19] Securities Registration Statement (S-1) under the Securities Act of 1933, 18 August 2004.

[20] Mark Zuckerberg, "Building global community", Facebook, 16 February 2017.

[21] Twitter, S-1 Registration Statement, 13 October 2013, pp. 91–92.

[22] VKontakte, company information; Tencent, "About Tencent".

[23] Danish Institute for Human Rights submission. Cf. Yahoo/Oath submission, 2016.

[24] Guiding Principles, principle 11.

(e) Conduct ongoing review of their efforts to respect rights, including through regular consultation with stakeholders, and frequent, accessible and effective communication with affected groups and the public (principles 20–21);

(f) Provide appropriate remediation, including through operational-level grievance mechanisms that users may access without aggravating their "sense of disempowerment" (principles 22, 29 and 31).

III. Key concerns with content regulation

12. Governments seek to shape the environment in which companies moderate content, while the companies predicate individual access to their platforms on user agreement with terms of service that govern what may be expressed and how individuals may express it.

A. Government regulation

13. States regularly require companies to restrict manifestly illegal content such as representations of child sexual abuse, direct and credible threats of harm and incitement to violence, presuming they also meet the conditions of legality and necessity.[25] Some States go much further and rely on censorship and criminalization to shape the online regulatory environment.[26] Broadly worded restrictive laws on "extremism", blasphemy, defamation, "offensive" speech, "false news" and "propaganda" often serve as pretexts for demanding that companies suppress legitimate discourse.[27] Increasingly, States target content specifically on online platforms.[28] Other laws may interfere with online privacy in ways that deter the exercise of freedom of opinion and expression.[29] Many States also deploy tools of disinformation and propaganda to limit the accessibility and trustworthiness of independent media.[30]

14. *Liability protections*. From early in the digital age, many States adopted rules to protect intermediaries from liability for the content third parties publish on their platforms. The European Union e-commerce directive, for instance, establishes a legal regime to protect intermediaries from liability for content except when they go beyond their role as a "mere conduit", "cache" or "host" of information provided by users.[31] Section 230 of the United States Communications Decency Act generally provides immunity for providers of "interactive computer service[s]" that host or publish information about others, but this has since been curtailed.[32] The intermediary liability regime in Brazil requires a court order to restrict particular content,[33] while the intermediary liability regime in India establishes a "notice and takedown" process that involves the order of a court or similar adjudicative

[25] Ireland has established co-regulatory mechanisms with companies to restrict illegal child sexual abuse material: Ireland submission. Many companies rely on a picture recognition algorithm to detect and remove child pornography: submissions by Open Technology Institute, p. 2 and ARTICLE 19, p. 8.

[26] See A/HRC/32/38, paras. 46–47. On Internet shutdowns, see A/HRC/35/22, paras. 8–16 and examples of communications of the Special Rapporteur: Nos. UA TGO 1/2017, UA IND 7/2017 and AL GMB 1/2017.

[27] Communication Nos. OL MYS 1/2018, UA RUS 7/2017, UA ARE 7/2017, AL BHR 8/2016, AL SGP 5/2016 and OL RUS 7/2016. Azerbaijan prohibits propaganda of terrorism, religious extremism and suicide: Azerbaijan submission.

[28] See communication Nos. OL PAK 8/2016 and OL LAO 1/2014; Association for Progressive Communications, *Unshackling Expression: A Study on Laws Criminalising Expression Online in Asia*, GISWatch 2017 Special Edition.

[29] A/HRC/29/32.

[30] See, for example, Gary King, Jennifer Pan and Margaret E. Roberts, "How the Chinese Government fabricates social media posts for strategic distraction, not engaged argument", *American Political Science Review*, vol. 111, No. 3 (2017), pp. 484–501.

[31] Directive No. 2000/31/EC of the European Parliament and of the Council of 8 June 2000.

[32] 47 United States Code § 230. See also the Allow States and Victims to Fight Online Sex Trafficking Act (H.R. 1865).

[33] *Marco Civil da Internet*, federal law 12.965, arts. 18–19.

body.[34] The 2014 Manila Principles on Intermediary Liability, developed by a coalition of civil society experts, identify essential principles that should guide any intermediary liability framework.

15. *Imposition of company obligations*. Some States impose obligations on companies to restrict content under vague or complex legal criteria without prior judicial review and with the threat of harsh penalties. For example, the Chinese Cybersecurity Law of 2016 reinforces vague prohibitions against the spread of "false" information that disrupts "social or economic order", national unity or national security; it also requires companies to monitor their networks and report violations to the authorities.[35] Failure to comply has reportedly led to heavy fines for the country's biggest social media platforms.[36]

16. Obligations to monitor and rapidly remove user-generated content have also increased globally, establishing punitive frameworks likely to undermine freedom of expression even in democratic societies. The network enforcement law (*NetzDG*) in Germany requires large social media companies to remove content inconsistent with specified local laws, with substantial penalties for non-compliance within very short time frames.[37] The European Commission has even recommended that member States establish legal obligations for active monitoring and filtering of illegal content.[38] Guidelines adopted in 2017 in Kenya on the dissemination of social media content during elections require platforms to "pull down accounts used in disseminating undesirable political contents on their platforms" within 24 hours.[39]

17. In the light of legitimate State concerns such as privacy and national security, the appeal of regulation is understandable. However, such rules involve risks to freedom of expression, putting significant pressure on companies such that they may remove lawful content in a broad effort to avoid liability. They also involve the delegation of regulatory functions to private actors that lack basic tools of accountability. Demands for quick, automatic removals risk new forms of prior restraint that already threaten creative endeavours in the context of copyright.[40] Complex questions of fact and law should generally be adjudicated by public institutions, not private actors whose current processes may be inconsistent with due process standards and whose motives are principally economic.[41]

18. *Global removals*. Some States are demanding extraterritorial removal of links, websites and other content alleged to violate local law.[42] Such demands raise serious

---

[34] Supreme Court of India, *Shreya Singhal v. Union of India*, decision of 24 March 2015.

[35] Articles 12 and 47, Human Rights in China submission, 2016, p. 12. For comments on an earlier draft of the Cybersecurity Law, see communication No. OL CHN 7/2015. See also Global Voices, "Netizen Report: Internet censorship bill looms large over Egypt", 16 March 2018; Republic of South Africa, Films and Publications Amendment Bill (B 61—2003).

[36] PEN America, *Forbidden Feeds: Government Controls on Social Media in China* (2018), p. 21.

[37] Act to Improve Enforcement of the Law in Social Networks (Network Enforcement Act), July 2017. See communication No. OL DEU 1/2017.

[38] European Commission, recommendation on measures to effectively tackle illegal content online (last updated: 5 March 2018).

[39] See communication No. OL KEN 10/2017; Javier Pallero, "Honduras' new bill threatens to curb online speech", Access Now, 12 February 2018.

[40] See European Commission, Proposal for a Directive of the European Parliament and of the Council on Copyright in the Digital Single Market, COM (2016) 593 final, art. 13; Daphne Keller, "Problems with filters in the European Commission's platforms proposal", Stanford Law School Center for Internet and Society, 5 October 2017; Fundación Karisma submission, 2016, pp. 4–6.

[41] Under European Union law, search engines are required to determine the validity of claims brought under the "right to be forgotten" framework. European Court of Justice, *Google Spain v. Agencia Española de Protección de Datos and Mario Costeja González* (case C-131/12), judgment (Grand Chamber) of 13 May 2014; submissions by ARTICLE 19, pp. 2–3 and Access Now, pp. 6–7; Google, "Updating our 'right to be forgotten' Transparency Report"; Theo Bertram and others, *Three Years of the Right to be Forgotten* (Google, 2018).

[42] See, for example, PEN America, *Forbidden Feeds*, pp. 36–37; Supreme Court of Canada, *Google Inc. v. Equustek Solutions Inc.*, judgment of 28 June 2017; European Court of Justice, *Google Inc. v.*

concern that States may interfere with the right to freedom of expression "regardless of frontiers". The logic of these demands would allow censorship across borders, to the benefit of the most restrictive censors. Those seeking removals should be required to make such requests in every jurisdiction where relevant, through regular legal and judicial process.

19. *Government demands not based on national law*. Companies distinguish between requests for the removal of allegedly illegal content submitted through regular legal channels and requests for removal based on the companies' terms of service.[43] (Legal removals generally apply only in the requesting jurisdiction; terms of service removals generally apply globally.) State authorities increasingly seek content removals outside of legal process or even through terms of service requests.[44] Several have established specialized government units to refer content to companies for removal. The European Union Internet Referral Unit, for instance, "flag[s] terrorist and violent extremist content online and cooperat[es] with online service providers with the aim of removing this content".[45] Australia also has similar referral mechanisms.[46] In South-East Asia, parties allied with Governments reportedly attempt to use terms of service requests to restrict political criticism.[47]

20. States also place pressure on companies to accelerate content removals through non-binding efforts, most of which have limited transparency. A three-year ban on YouTube in Pakistan compelled Google to establish a local version susceptible to government demands for removals of "offensive" content.[48] Facebook and Israel reportedly agreed to coordinate efforts and staff to monitor and remove "incitement" online. The details of this agreement were not disclosed, but the Israeli Minister of Justice claimed that between June and September 2016, Facebook granted nearly all government requests for removal of "incitement".[49] Arrangements to coordinate content actions with State input exacerbate concerns that companies perform public functions without the oversight of courts and other accountability mechanisms.[50]

21. The 2016 European Union Code of Conduct on countering illegal hate speech online involves agreement between the European Union and four major companies to remove content, committing them to collaborate with "trusted flaggers" and promote "independent counter-narratives".[51] While the promotion of counter-narratives may be attractive in the face of "extremist" or "terrorist" content, pressure for such approaches runs the risk of transforming platforms into carriers of propaganda well beyond established areas of legitimate concern.[52]

Commission nationale de l'informatique et des libertés (CNIL) (case C-507/17), Global Network Initiative submission, p. 6.

[43] Compare Twitter Transparency Report: Removal Requests (January–June 2017) with Twitter Transparency Report: Government Terms of Service Reports (January–June 2017). See also Facebook, Government requests: Frequently Asked Questions (FAQs).

[44] Submissions by ARTICLE 19, p. 2 and Global Network Initiative, p. 5.

[45] European Union, Internet Referral Unit, Year One Report, sect. 4.11; submissions by European Digital Rights (EDRi), p. 1 and Access Now, pp. 2–3.

[46] Australia submission.

[47] Southeast Asian Press Alliance, p. 1.

[48] Digital Rights Foundation submission.

[49] 7amleh – The Arab Center for the Advancement of Social Media submission.

[50] Association for Progressive Communications, p. 14 and 7amleh.

[51] "Trusted flaggers … refers to the status given to certain organisations which allows them to report illegal content through a special reporting system or channel, which is not available to normal users." European Commission, Code of Conduct on countering illegal hate speech online: First results on implementation (December 2016).

[52] The same companies created the Global Internet Forum to Counter Terrorism, an effort to develop industry-wide technological tools to remove terrorist content on their platforms. Google, "Update on the Global Internet Forum to Counter Terrorism", 4 December 2017.

B. Company moderation of content

Company compliance with national law

22. Each company is committed in principle to comply with the local law where it does business. As Facebook puts it: "If, after careful legal review, we determine that the content is illegal under local law, then we make it unavailable in the relevant country or territory."[53] Tencent, the owner of the mobile chat and social media app WeChat, goes considerably further, requiring anyone using the platform within China and Chinese citizens using the platform "anywhere in the world" to comply with content restrictions that mirror Chinese law or policy.[54] Several companies also collaborate with one another and regulatory bodies to remove images of child sexual abuse.[55]

23. The commitment to legal compliance can be complicated when relevant State law is vague, subject to varying interpretations or inconsistent with human rights law. For instance, laws against "extremism" which leave the key term undefined provide discretion to government authorities to pressure companies to remove content on questionable grounds.[56] Similarly, companies are often under pressure to comply with State laws that criminalize content that is said to be, for instance, blasphemous, critical of the State, defamatory of public officials or false. As explained below, the Guiding Principles provide tools to minimize the impact of such laws on individual users. The Global Network Initiative, a multi-stakeholder initiative that helps ICT companies navigate human rights challenges, has developed additional guidance on how to employ these tools.[57] One tool of minimization is transparency: many companies report annually on the number of government requests they receive and execute per State.[58] However, companies do not consistently disclose sufficient information about how they respond to government requests, nor do they regularly report government requests made under terms of service.[59]

Company moderation standards

24. Internet companies require their users to abide by terms of service and "community standards" that govern expression on their platforms.[60] Company terms of service, which users are required to accept in exchange for use of the platform, identify jurisdictions for dispute resolution and reserve to themselves discretion over content and account actions.[61] Platform content policies are a subset of these terms, articulating constraints on what users may express and how they may express it. Most companies do not explicitly base content

---

[53] Facebook, Government requests FAQs. See also Google legal removal requests, Twitter rules and policies, Reddit content policy.

[54] Tencent, Terms of Service: Introduction, Tencent, Agreement on Software License and Service of Tencent Wenxin.

[55] United Nations Educational, Scientific and Cultural Organization, *Fostering Freedom Online: The Role of Internet Intermediaries* (Paris, 2014), pp. 56–57.

[56] See Maria Kravchenko, "Inappropriate enforcement of anti-extremist legislation in Russia in 2016", SOVA Center for Information and Analysis, 21 April 2017; Danielle Citron, "Extremist speech, compelled conformity, and censorship creep", *Notre Dame Law Review*, vol. 93, No. 3 (2018), pp. 1035–1071.

[57] Global Network Initiative, Principles on Freedom of Expression and Privacy, sect. 2. Social media companies participating in the Initiative include Facebook, Google, Microsoft/LinkedIn and Yahoo/Oath.

[58] See paragraph 39 below. In addition, Automattic, Google, Microsoft/Bing and Twitter are among the companies that regularly, although not necessarily comprehensively, post government takedown and intellectual property requests to the Lumen database.

[59] Ranking Digital Rights, 2017 Corporate Accountability Index, p. 28.

[60] Jamila Venturini and others, *Terms of Service and Human Rights: An Analysis of Online Platform Contracts* (Rio de Janeiro, Revan, 2016).

[61] Baidu user agreement ("[We] remove and delete any content in this service based on Baidu's own discretion for any reason."); Tencent terms of service ("We reserve the right to block or remove Your Content for any reason, including as is in our opinion appropriate or as required by applicable laws and regulations."); Twitter terms of service ("We may suspend or terminate your account or cease providing you with all or part of the Services at any time for any or no reason.").

standards on any particular body of law that might govern expression, such as national law or international human rights law. The Chinese search giant Baidu, however, prohibits content that is "opposed to the basic principles established by the Constitution" of the People's Republic of China.[62]

25. The development of content moderation policies typically involves legal counsel, public policy and product managers, and senior executives. Companies may establish "trust and safety" teams to address spam, fraud and abuse, and counter-terrorism teams may address terrorist content.[63] Some have developed mechanisms for soliciting input from outside groups on specialized aspects of content policies.[64] The exponential increase in user-generated content has triggered the development of detailed and constantly evolving rules. These rules vary according to a range of factors, from company size, revenue and business model to the "platform's brand and reputation, its tolerance for risk, and the type of user engagement it wishes to attract".[65]

Areas of concern around content standards

26. *Vague rules*. Company prohibitions of threatening or promoting terrorism,[66] supporting or praising leaders of dangerous organizations[67] and content that promotes terrorist acts or incites violence[68] are, like counter-terrorism legislation, excessively vague.[69] Company policies on hate, harassment and abuse also do not clearly indicate what constitutes an offence. Twitter's prohibition of "behavior that incites fear about a protected group" and Facebook's distinction between "direct attacks" on protected characteristics and merely "distasteful or offensive content" are subjective and unstable bases for content moderation.[70]

27. *Hate, harassment, abuse*. The vagueness of hate speech and harassment policies has triggered complaints of inconsistent policy enforcement that penalizes minorities while reinforcing the status of dominant or powerful groups. Users and civil society report violence and abuse against women, including physical threats, misogynist comments, the posting of non-consensual or fake intimate images and doxing;[71] threats of harm against the politically disenfranchised,[72] minority races and castes[73] and ethnic groups suffering from violent persecution;[74] and abuse directed at refugees, migrants and asylum seekers.[75] At the same time, platforms have reportedly suppressed lesbian, gay, bisexual, transgender and queer activism,[76] advocacy against repressive Governments,[77] reporting on ethnic cleansing[78] and critiques of racist phenomena and power structures.[79]

---

[62] Baidu terms of service, sect. 3.1.

[63] Monika Bickert, "Hard questions: how we counter terrorism", 15 June 2017.

[64] See, for example, Twitter Trust and Safety Council and YouTube Trusted Flagger Program.

[65] Sarah Roberts, *Content Moderation* (University of California at Los Angeles, 2017). See also ARTICLE 19 submission, p. 2.

[66] Twitter rules and policies (violent extremist groups).

[67] Facebook community standards (dangerous organizations).

[68] YouTube policies (violent or graphic content policies).

[69] See A/HRC/31/65, para. 39.

[70] Facebook community standards (hate speech); Twitter rules and policies (hateful conduct policy).

[71] Amnesty International, *Toxic Twitter: A Toxic Place for Women*; Association for Progressive Communications submission, p. 2.

[72] Submissions by 7amleh and Association for Progressive Communications, p. 15.

[73] Ijeoma Oluo, "Facebook's complicity in the silencing of black women", Medium, 2 August 2017; submissions by Center for Communications Governance, p. 5 and Association for Progressive Communications, pp. 11–12.

[74] Statement by the Special Rapporteur on the situation of human rights in Myanmar, Yanghee Lee, to the thirty-seventh session of the Human Rights Council, 12 March 2018.

[75] Association for Progressive Communications submission, p. 12.

[76] Electronic Frontier Foundation submission, p. 5.

[77] Ibid.; submissions by Association for Progressive Communications and 7amleh.

[78] Betsy Woodruff, "Facebook silences Rohingya reports of ethnic cleansing", The Daily Beast, 18 September 2017; ARTICLE 19 submission, p. 9.

28. The scale and complexity of addressing hateful expression presents long-term challenges and may lead companies to restrict such expression even if it is not clearly linked to adverse outcomes (as hateful advocacy is connected to incitement in article 20 of the International Covenant on Civil and Political Rights). Companies should articulate the bases for such restrictions, however, and demonstrate the necessity and proportionality of any content actions (such as removals or account suspensions). Meaningful and consistent transparency about enforcement of hate speech policies, through substantial reporting of specific cases, may also provide a level of insight that even the most detailed explanations cannot offer.[80]

29. *Context*. Companies emphasize the importance of context when assessing the applicability of general restrictions.[81] Nonetheless, attention to context has not prevented removals of depictions of nudity with historical, cultural or educational value;[82] historical and documentary accounts of conflict;[83] evidence of war crimes;[84] counter speech against hate groups;[85] or efforts to challenge or reclaim racist, homophobic or xenophobic language.[86] Meaningful examination of context may be thwarted by time and resource constraints on human moderators, overdependence on automation or insufficient understanding of linguistic and cultural nuance.[87] Companies have urged users to supplement controversial content with contextual details, but the feasibility and effectiveness of this guidance are unclear.[88]

30. *Real-name requirements*. In order to deal with online abuse, some companies have "authentic identity" requirements;[89] others approach identity questions more flexibly.[90] The effectiveness of real-name requirements as safeguards against online abuse is questionable.[91] Indeed, strict insistence on real names has unmasked bloggers and activists using pseudonyms to protect themselves, exposing them to grave physical danger.[92] It has also blocked the accounts of lesbian, gay, bisexual, transgender and queer users and activists, drag performers and users with non-English or unconventional names.[93] Since online anonymity is often necessary for the physical safety of vulnerable users, human rights principles default to the protection of anonymity, subject only to limitations that would protect their identities.[94] Narrowly crafted impersonation rules that limit the ability of users to portray another person in a confusing or deceptive manner may be a more proportionate means of protecting the identity, rights and reputations of other users.[95]

31. *Disinformation*. Disinformation and propaganda challenge access to information and the overall public trust in media and government institutions. The companies face

---

[79] Julia Angwin and Hannes Grassegger, "Facebook's secret censorship rules protect white men from hate speech but not black children", ProPublica, 28 June 2017.

[80] See paras. 52 and 62 below.

[81] Twitter, "Our approach to policy development and enforcement philosophy"; YouTube policies (the importance of context); Richard Allan, "Hard questions: who should decide what is hate speech in an online global community?", Facebook Newsroom, 27 June 2017.

[82] Submissions by OBSERVACOM, p. 11 and ARTICLE 19, p. 6.

[83] WITNESS submission, pp. 6–7.

[84] Ibid.

[85] Electronic Frontier Foundation submission, p. 5.

[86] Association for Progressive Communications submission, p. 14.

[87] See Allan, "Hard questions".

[88] YouTube policies (the importance of context); Facebook community standards (hate speech).

[89] Facebook community standards (using your authentic identity). Note that Facebook now permits exceptions to its real-name policy on a case-by-case basis, but this has been criticized as insufficient: Access Now submission, p. 12. Baidu even requires the use of personally identifying information: Baidu user agreement.

[90] Twitter Help Center, "Help with username registration"; Instagram, "Getting started on Instagram".

[91] J. Nathan Matias, "The real name fallacy", Coral Project, 3 January 2017.

[92] Access Now submission, p. 11.

[93] Dia Kayyali, "Facebook's name policy strikes again, this time at Native Americans", Electronic Frontier Foundation, 13 February 2015.

[94] See A/HRC/29/32, para. 9.

[95] Twitter rules and policies (impersonation policy).

increasing pressure to address disinformation spread through links to bogus third-party news articles or websites, fake accounts, deceptive advertisements and the manipulation of search rankings.[96] However, because blunt forms of action, such as website blocking or specific removals, risk serious interference with freedom of expression, companies should carefully craft any policies dealing with disinformation.[97] Companies have adopted a variety of responses, including arrangements with third-party fact checkers, heightened enforcement of advertisement policies, enhanced monitoring of suspicious accounts, changes in content curation and search ranking algorithms, and user trainings on identifying false information.[98] Some measures, particularly those that enhance restrictions on news content, may threaten independent and alternative news sources or satirical content.[99] Government authorities have taken positions that may reflect outsized expectations about technology's power to solve such problems alone.[100]

Company moderation processes and tools

32. *Automated flagging, removal and pre-publication filtering*. The massive scale of user-generated content has led the largest companies to develop automated moderation tools. Automation has been employed primarily to flag content for human review, and sometimes to remove it. Automated tools scanning music and video for copyright infringement at the point of upload have raised concerns of overblocking, and calls to expand upload filtering to terrorist-related and other areas of content threaten to establish comprehensive and disproportionate regimes of pre-publication censorship.[101]

33. Automation may provide value for companies assessing huge volumes of user-generated content, with tools ranging from keyword filters and spam detection to hash-matching algorithms and natural language processing.[102] Hash matching is widely used to identify child sexual abuse images, but its application to "extremist" content — which typically requires assessment of context — is difficult to accomplish without clear rules regarding "extremism" or human evaluation.[103] The same is true with natural language processing.[104]

34. *User and trusted flagging*. User flags give individuals the ability to log complaints of inappropriate content with content moderators. Flags typically do not enable nuanced discussions about appropriate boundaries (e.g., why content may be offensive but, on balance, better left up).[105] They have also been "gamed" to heighten pressure on platforms to remove content supportive of sexual minorities and Muslims.[106] Many companies have developed specialized rosters of "trusted" flaggers, typically experts, high-impact users and, reportedly, sometimes government flaggers.[107] There is little or no public information

---

[96] Ibid., Allen Babajanian and Christine Wendel, "#FakeNews: innocuous or intolerable?", Wilton Park report 1542, April 2017.

[97] Joint Declaration 2017.

[98] Submissions by Association for Progressive Communications, pp. 4–6 and ARTICLE 19, p. 4.

[99] Association for Progressive Communications submission, p. 5.

[100] See communication No. OL ITA 1/2018. Cf. European Commission, *A Multi-Dimensional Approach to Disinformation: Final Report of the Independent High-level Group on Fake News and Disinformation* (Luxembourg, 2018).

[101] The United Kingdom of Great Britain and Northern Ireland reportedly developed a tool to automatically detect and remove terrorist content at the point of upload. Home Office, "New technology revealed to help fight terrorist content online", 13 February 2018.

[102] Center for Democracy and Technology, *Mixed Messages? The Limits of Automated Media Content Analysis* (November 2017), p. 9.

[103] Open Technology Institute submission, p. 2.

[104] Center for Democracy and Technology, *Mixed Messages?*, p. 4.

[105] On user flags, see generally Kate Crawford and Tarleton Gillespie, "What is a flag for? Social media reporting tools and the vocabulary of complaint", *New Media and Society*, vol. 18, No. 3 (March 2016), pp. 410–428.

[106] Ibid., p. 421.

[107] YouTube Help, YouTube Trusted Flagger Program, YouTube Help, "Get involved with YouTube contributors".

explaining the selection of specialized flaggers, their interpretations of legal or community standards or their influence over company decisions.

35. *Human evaluation.* Automation often will be supplemented by human review, with the biggest social media companies developing large teams of content moderators to review flagged content.[108] Flagged content may be routed to content moderators, which will typically be authorized to make a decision — often within minutes — about the appropriateness of the content and to remove or permit it. In situations where the appropriateness of particular content is difficult to determine, moderators may escalate its review to content teams at company headquarters. In turn, company officials — typically public policy or "trust and safety" teams with the engagement of general counsel — will make decisions on removals. Company disclosure about removal discussions, in aggregate or specific cases, is limited.[109]

36. *Account or content action.* The existence of inappropriate content may trigger a range of company actions. Companies may limit content removal by jurisdiction, a range of jurisdictions, or across an entire platform or set of platforms. They may apply age limitations, warnings or demonetization.[110] Violations may lead to temporary account suspensions, while repeat offences may lead to account deactivation. In very few cases outside of copyright enforcement do the companies provide "counter-notice" procedures that permit users posting content to challenge removals.

37. *Notification.* A common complaint is that users who post reported content, or persons complaining of abuse, may not receive any notification of removal or other action.[111] Even when companies issue notifications, these typically indicate merely the action taken and a generic ground for action. At least one company has attempted to provide more context in its notifications, but it is unclear whether additional detail in stock notifications constitutes sufficient explanation in all cases.[112] Transparency and notifications go hand in hand: robust operational-level transparency that improves user awareness of the platform's approaches to content removals alleviates the pressure on notifications in individual cases, while weaker overall transparency increases the likelihood that users will be unable to understand individual removals in the absence of notifications tailored to specific cases.

38. *Appeals and remedies.* Platforms permit appeals of a range of actions, from profile or page removals to removals of specific posts, photos or videos.[113] Even with appeal, however, the remedies available to users appear limited or untimely to the point of non-existence and, in any event, opaque to most users and even civil society experts. It may be, for instance, that reinstatement of content would be an insufficient response if removal resulted in specific harm — such as reputational, physical, moral or financial — to the person posting. Similarly, account suspensions or content removals during public protest or debate could have significant impact on political rights and yet lack any company remedy.

Transparency

39. Companies have developed transparency reports that publish aggregated data on government requests for content removal and user data. Such reporting demonstrates the kinds of pressures the companies face. Transparency reporting identifies, country by

[108] See Sarah Roberts, "Commercial content moderation: digital laborers' dirty work", *Media Studies Publications*, paper 12 (2016).

[109] Cf. Wikipedia: BOLD, revert, discuss cycle. Reddit moderators are encouraged to offer "helpful rule explanations, tips and links to new and confused users" (Reddit Moddiquette).

[110] YouTube policies (nudity and sexual content policies); YouTube Help, "Creator influence on YouTube".

[111] Submissions by ARTICLE 19, p. 7 and Association for Progressive Communications, p. 16.

[112] See https://twitter.com/TwitterSafety/status/971882517698510848/.

[113] Electronic Frontier Foundation and Visualizing Impact, "How to appeal", onlinecensorship.org. Facebook and Instagram allow only the appeal of account suspensions. Cf. Github submission, p. 6.

country, the number of legal removal requests,[114] the number of requests where some action was taken or content restricted[115] and, increasingly, descriptions and examples of selected legal bases.[116]

40. However, as the leading review of Internet transparency concludes, companies disclose "the least amount of information about how *private* rules and mechanisms for self- and co-regulation are formulated and carried out".[117] In particular, disclosure concerning actions taken pursuant to private removal requests under terms of service is "incredibly low".[118] Content standards are drafted in broad terms, leaving room for platform discretion that companies do not sufficiently illuminate. Media and public scrutiny have led companies to supplement general policies with explanatory blog posts[119] and limited hypothetical examples,[120] but these fall short of illuminating nuances in how internal rules are developed and applied.[121] While terms of service are generally available in local languages, transparency reports, company blogs and related content are not, providing even less clarity to non-English-speaking users. Accordingly, users, public authorities and civil society often express dissatisfaction with the unpredictability of terms of service actions.[122] The lack of sufficient engagement, coupled with growing public criticism, has forced companies into a constant state of rule evaluation, revision and defence.

IV. Human rights principles for company content moderation

41. The founder of Facebook recently expressed his hope for a process in which the company "could more accurately reflect the values of the community in different places".[123] That process, and the relevant standards, can be found in human rights law. Private norms, which vary according to each company's business model and vague assertions of community interests, have created unstable, unpredictable and unsafe environments for users and intensified government scrutiny. National laws are inappropriate for companies that seek common norms for their geographically and culturally diverse user base. But human rights standards, if implemented transparently and consistently with meaningful user and civil society input, provide a framework for holding both States and companies accountable to users across national borders.

42. A human rights framework enables forceful normative responses against undue State restrictions — provided companies play by similar rules. The Guiding Principles and their accompanying body of "soft law" provide guidance on how companies should prevent or mitigate government demands for excessive content removals. But they also establish principles of due diligence, transparency, accountability and remediation that limit platform interference with human rights through product and policy development. Companies committed to implementing human rights standards throughout their operations — and not merely when it aligns with their interests — will stand on firmer ground when they seek to

[114] Twitter Transparency Report: Removal Requests (January–June 2017); Google Transparency Report: Government Requests to Remove Content; 2016 Reddit Inc., Transparency Report. Facebook does not provide the total number of requests received per country.

[115] See, for example, Facebook Transparency Report (France) (January–June 2017); Google Transparency Report: Government Requests to Remove Content (India); Twitter Transparency Report (Turkey).

[116] Ibid.

[117] Ranking Digital Rights submission, p. 4. Original italics.

[118] Ibid., p. 10.

[119] See Elliot Schrage, "Introducing hard questions", Facebook Newsroom, 15 June 2017; Twitter Safety, "Enforcing new rules to reduce hateful conduct and abusive behavior", 18 December 2017.

[120] See, for example, YouTube policies (violent or graphic content policies).

[121] Angwin and Grasseger, "Facebook's secret censorship rules".

[122] Submissions by Ranking Digital Rights, p. 10; OBSERVACOM p. 10; Association for Progressive Communications, p. 17; International Federation of Library Associations and Institutions, pp. 4–5; Access Now, p. 17; and EDRi, p. 5.

[123] Kara Swisher and Kurt Wagner, "Here's the transcript of Recode's interview with Facebook CEO Mark Zuckerberg about the Cambridge Analytica controversy and more", Recode, 22 March 2018.

hold States accountable to the same standards. Furthermore, when companies align their terms of service more closely with human rights law, States will find it harder to exploit them to censor content.

43. Human rights principles also enable companies to create an inclusive environment that accommodates the varied needs and interests of their users while establishing predictable and consistent baseline standards of behaviour. Amidst growing debate about whether companies exercise a combination of intermediary and editorial functions, human rights law expresses a promise to users that they can rely on fundamental norms to protect their expression over and above what national law might curtail.[124] Yet human rights law is not so inflexible or dogmatic that it requires companies to permit expression that would undermine the rights of others or the ability of States to protect legitimate national security or public order interests. Across a range of ills that may have more pronounced impact in digital space than they might offline — such as misogynist or homophobic harassment designed to silence women and sexual minorities, or incitement to violence of all sorts — human rights law would not deprive companies of tools. To the contrary, it would offer a globally recognized framework for designing those tools and a common vocabulary for explaining their nature, purpose and application to users and States.

A. Substantive standards for content moderation

44. The digital age enables rapid dissemination and enormous reach, but it also lacks textures of human context. Per the Guiding Principles, companies may take into account the size, structure and distinctive functions of the platforms they provide in assessing the necessity and proportionality of content restrictions.

45. *Human rights by default*. Terms of service should move away from a discretionary approach rooted in generic and self-serving "community" needs. Companies should instead adopt high-level policy commitments to maintain platforms for users to develop opinions, express themselves freely and access information of all kinds in a manner consistent with human rights law.[125] These commitments should govern their approach to content moderation and to complex problems such as computational propaganda[126] and the collection and handling of user data. Companies should incorporate directly into their terms of service and "community standards" relevant principles of human rights law that ensure content-related actions will be guided by the same standards of legality, necessity and legitimacy that bind State regulation of expression.[127]

46. "*Legality*". Company rules routinely lack the clarity and specificity that would enable users to predict with reasonable certainty what content places them on the wrong side of the line. This is particularly evident in the context of "extremism" and hate speech, areas of restriction easily susceptible to excessive removals in the absence of rigorous human evaluation of context. Further complicating public understanding of context-specific rules is the emerging general exception for "newsworthiness".[128] While the recognition of public interest is welcome, companies should also explain what factors are assessed in determining the public interest and what factors other than public interest inform calculations of newsworthiness. Companies should supplement their efforts to explain their rules in more detail with aggregate data illustrating trends in rule enforcement, and examples of actual cases or extensive, detailed hypotheticals that illustrate the nuances of interpretation and application of specific rules.

47. *Necessity and proportionality*. Companies should not only describe contentious and context-specific rules in more detail. They should also disclose data and examples that

[124] Global Partners Digital submission, p. 3; Guiding Principles, principle 11.

[125] Guiding Principles, principle 16.

[126] See Samuel Wooley and Philip Howard, *Computational Propaganda Worldwide: Executive Summary* (Computational Propaganda Research Project working paper No. 2017.11 (Oxford, 2017).

[127] Global Partners Digital submission, pp. 10–13.

[128] See Joel Kaplan, "Input from community and partners on our community standards", Facebook Newsroom, 21 October 2016; Twitter rules and policies.

provide insight into the factors they assess in determining a violation, its severity and the action taken in response. In the context of hate speech, explaining how specific cases are resolved may help users better understand how companies approach difficult distinctions between offensive content and incitement to hatred, or how considerations such as the intent of the speaker or the likelihood of violence are assessed in online contexts. Granular data on actions taken will also establish a basis to evaluate the extent to which companies are narrowly tailoring restrictions. The circumstances under which they apply less intrusive restrictions (such as warnings, age restrictions or demonetization) should be explained.

48. *Non-discrimination.* Meaningful guarantees of non-discrimination require companies to transcend formalistic approaches that treat all protected characteristics as equally vulnerable to abuse, harassment and other forms of censorship.[129] Indeed, such approaches would appear inconsistent with their own emphasis that context matters. Instead, when companies develop or modify policies or products, they should actively seek and take into account the concerns of communities historically at risk of censorship and discrimination.

B. Processes for company moderation and related activities

Responses to government requests

49. As company transparency reports show, Governments pressure them to remove content, suspend accounts and identify and disclose account information. Where required by local law, it may appear that companies have little choice but to comply. But companies may develop tools that prevent or mitigate the human rights risks caused by national laws or demands inconsistent with international standards.

50. *Prevention and mitigation.* Companies often claim to take human rights seriously. But it is not enough for companies to undertake such commitments internally and provide ad hoc assurances to the public when controversies arise. Companies should also, at the highest levels of leadership, adopt and then publicly disclose specific policies that "direct all business units, including local subsidiaries, to resolve any legal ambiguity in favour of respect for freedom of expression, privacy, and other human rights". Policies and procedures that interpret and implement government demands to narrow and "ensure the least restriction on content" should flow from these commitments.[130] Companies should ensure that requests are in writing, cite specific and valid legal bases for restrictions and are issued by a valid government authority in an appropriate format.[131]

51. When faced with problematic requests, companies should seek clarification or modification; solicit the assistance of civil society, peer companies, relevant government authorities, international and regional bodies and other stakeholders; and explore all legal options for challenge.[132] When companies receive requests from States under their terms of service or through other extralegal means, they should route these requests through legal compliance processes and assess the validity of such requests under relevant local laws and human rights standards.

52. *Transparency.* In the face of censorship and associated human rights risks, users can only make informed decisions about whether and how to engage on social media if interactions between companies and States are meaningfully transparent. Best practices on how to provide such transparency should be developed. Company reporting about State requests should be supplemented with granular data concerning the types of requests received (e.g., defamation, hate speech, terrorism-related content) and actions taken (e.g., partial or full removal, country-specific or global removal, account suspension, removal granted under terms of service). Companies should also provide specific examples as often

[129] See, for example, International Convention on the Elimination of All Forms of Racial Discrimination, arts. 1 (4) and 2 (2).

[130] See A/HRC/35/22, paras. 66–67.

[131] Submissions by Global Network Initiative, pp. 3–4 and GitHub, pp. 3–5.

[132] See A/HRC/35/22, para. 68.

as possible.[133] Transparency reporting should extend to government demands under company terms of service[134] and must also account for public-private initiatives to restrict content, such as the European Union Code of Conduct on countering illegal hate speech online, governmental initiatives such as Internet referral units and bilateral understandings such as those reported between YouTube and Pakistan and Facebook and Israel. Companies should preserve records of requests made under these initiatives and communications between the company and the requester and explore arrangements to submit copies of such requests to a third-party repository.

Rule-making and product development

53. *Due diligence*. Although several companies commit to human rights due diligence in assessing their response to State restrictions, it is unclear whether they implement similar safeguards to prevent or mitigate risks to freedom of expression posed by the development and enforcement of their own policies.[135] Companies should develop clear and specific criteria for identifying activities that trigger such assessments. In addition to revisions of content moderation policies and processes, assessments should be conducted on the curation of user feeds and other forms of content delivery, the introduction of new features or services and modifications to existing ones, the development of automation technologies and market-entry decisions such as arrangements to provide country-specific versions of the platform.[136] Past reporting also specifies the issues these assessments should examine and the internal processes and training required to integrate assessments and their findings into relevant operations. Additionally, these assessments should be ongoing and adaptive to changes in circumstances or operating context.[137] Multi-stakeholder initiatives such as Global Network Initiative provide an avenue for companies to develop and refine assessments and other due diligence processes.

54. *Public input and engagement*. Participants in consultations consistently raised concerns that companies failed to engage adequately with users and civil society, particularly in the global South. Input from affected rights holders (or their representatives) and relevant local or subject matter experts, and internal decision-making processes that meaningfully incorporate the feedback received, are integral components of due diligence.[138] Consultations — especially in broad forms such as calls for public comment — enable the companies to consider the human rights impact of their activities from diverse perspectives, while also encouraging them to pay close attention to how seemingly benign or ostensibly "community-friendly" rules may have significant, "hyper-local" impacts on communities.[139] For example, engagement with a geographically diverse range of indigenous groups may help companies develop better indicators for taking into account cultural and artistic context when assessing content featuring nudity.

55. *Rule-making transparency*. Companies too often appear to introduce products and rule modifications without conducting human rights due diligence or evaluating the impact in real cases. They should at least seek comment on their impact assessments from interested users and experts, in settings that guarantee the confidentiality of such assessments if necessary. They should also clearly communicate to the public the rules and processes that produced them.

---

[133] See, for example, Twitter Transparency Report: Removal Requests (January–June 2017).

[134] Twitter has begun to publish data on "non-legal requests submitted by known government representatives about content that may violate the Twitter Rules" prohibiting abusive behaviour, promotion of terrorism and intellectual property infringement. Ibid. See also Microsoft, Content Removal Requests Report (January–June 2017).

[135] Ranking Digital Rights submission, p. 12; Guiding Principles, principle 17.

[136] See A/HRC/35/22, para. 53.

[137] Ibid., paras. 54–58.

[138] See Guiding Principles, principle 18 and A/HRC/35/22, para. 57.

[139] Chinmayi Arun, "Rebalancing regulation of speech: hyper-local content on global web-based platforms", Berkman Klein Center for Internet and Society Medium Collection, Harvard University, 2018; *Pretoria News*, "Protest at Google, Facebook 'bullying' of bare-breasted maidens", 14 December 2017.

Rule enforcement

56. *Automation and human evaluation*. Automated content moderation, a function of the massive scale and scope of user-generated content, poses distinct risks of content actions that are inconsistent with human rights law. Company responsibilities to prevent and mitigate human rights impacts should take into account the significant limitations of automation, such as difficulties with addressing context, widespread variation of language cues and meaning and linguistic and cultural particularities. Automation derived from understandings developed within the home country of the company risks serious discrimination across global user bases. At a minimum, technology developed to deal with considerations of scale should be rigorously audited and developed with broad user and civil society input.

57. The responsibility to foster accurate and context-sensitive content moderation practices that respect freedom of expression also requires companies to strengthen and ensure professionalization of their human evaluation of flagged content. This strengthening should involve protections for human moderators consistent with human rights norms applicable to labour rights and a serious commitment to involve cultural, linguistic and other forms of expertise in every market where they operate. Company leadership and policy teams should also diversify to enable the application of local or subject-matter expertise to content issues.

58. *Notice and appeal*. Users and civil society experts commonly express concern about the limited information available to those subject to content removal or account suspension or deactivation, or those reporting abuse such as misogynistic harassment and doxing. The lack of information creates an environment of secretive norms, inconsistent with the standards of clarity, specificity and predictability. This interferes with the individual's ability to challenge content actions or follow up on content-related complaints; in practice, however, the lack of robust appeal mechanisms for content removals favours users who flag over those who post. Some may argue that it will be time-consuming and costly to allow appeals on every content action. But companies could work with one another and civil society to explore scalable solutions such as company-specific or industry-wide ombudsman programmes. Among the best ideas for such programmes is an independent "social media council", modelled on the press councils that enable industry-wide complaint mechanisms and the promotion of remedies for violations.[140] This mechanism could hear complaints from individual users that meet certain criteria and gather public feedback on recurrent content moderation problems such as overcensorship related to a particular subject area. States should be supportive of scalable appeal mechanisms that operate consistently with human rights standards.

59. *Remedy*. The Guiding Principles highlight the responsibility to remedy "adverse impacts" (principle 22). However, few if any of the companies provide for remediation. Companies should institute robust remediation programmes, which may range from reinstatement and acknowledgment to settlements related to reputational or other harms. There has been some convergence among several companies in their content rules, giving rise to the possibility of inter-company cooperation to provide remedies through a social media council, other ombudsman programmes or third-party adjudication. If the failure to remediate persists, legislative and judicial intervention may be required.

60. *User autonomy*. Companies have developed tools enabling users to shape their own online environments. This includes muting and blocking of other users or specific kinds of content. Similarly, platforms often permit users to create closed or private groups, moderated by users themselves. While content rules in closed groups should be consistent with baseline human rights standards, platforms should encourage such affinity-based groups given their value in protecting opinion, expanding space for vulnerable communities and allowing the testing of controversial or unpopular ideas. Real-name requirements

[140] See ARTICLE 19, *Self-regulation and 'Hate Speech' on Social Media Platforms* (London, 2018), pp. 20–22.

should be disfavoured, given their privacy and security implications for vulnerable individuals.[141]

61. Mounting concerns about the verifiability, relevance and usefulness of information online raise complex questions about how companies should respect the right to access information. At a minimum, companies should disclose details concerning their approaches to curation. If companies are ranking content on social media feeds based on interactions between users, they should explain the data collected about such interactions and how this informs the ranking criteria. Companies should provide all users with accessible and meaningful opportunities to opt out of platform-driven curation.[142]

Decisional transparency

62. Notwithstanding advances in aggregate transparency of government removal requests, terms of service actions are largely unreported. Companies do not publish data on the volume and type of private requests they receive under these terms, let alone rates of compliance. Companies should develop transparency initiatives that explain the impact of automation, human moderation and user or trusted flagging on terms of service actions. While a few companies are beginning to provide some information about these actions, the industry should be moving to provide more detail about specific and representative cases and significant developments in the interpretation and enforcement of their policies.

63. The companies are implementing "platform law", taking actions on content issues without significant disclosure about those actions. Ideally, companies should develop a kind of case law that would enable users, civil society and States to understand how the companies interpret and implement their standards. While such a "case law" system would not involve the kind of reporting the public expects from courts and administrative bodies, a detailed repository of cases and examples would clarify the rules much as case reporting does.[143] A social media council empowered to evaluate complaints across the ICT sector could be a credible and independent mechanism to develop such transparency.

V. Recommendations

64. **Opaque forces are shaping the ability of individuals worldwide to exercise their freedom of expression. This moment calls for radical transparency, meaningful accountability and a commitment to remedy in order to protect the ability of individuals to use online platforms as forums for free expression, access to information and engagement in public life. The present report has identified a range of steps, include the following.**

Recommendations for States

65. **States should repeal any law that criminalizes or unduly restricts expression, online or offline.**

66. **Smart regulation, not heavy-handed viewpoint-based regulation, should be the norm, focused on ensuring company transparency and remediation to enable the public to make choices about how and whether to engage in online forums. States should only seek to restrict content pursuant to an order by an independent and impartial judicial authority, and in accordance with due process and standards of legality, necessity and legitimacy. States should refrain from imposing disproportionate sanctions, whether heavy fines or imprisonment, on Internet intermediaries, given their significant chilling effect on freedom of expression.**

---

[141] See para. 30 above.

[142] Facebook, for example, permits users to view stories in their News Feed in reverse chronological order, but warns that it will "eventually" return to its default curation settings. Facebook Help Centre, "What's the difference between top stories and most recent stories on News Feed?"

[143] See, for example, Madeleine Varner and others, "What does Facebook consider hate speech?", ProPublica, 28 December 2017.

67. States and intergovernmental organizations should refrain from establishing laws or arrangements that would require the "proactive" monitoring or filtering of content, which is both inconsistent with the right to privacy and likely to amount to pre-publication censorship.

68. States should refrain from adopting models of regulation where government agencies, rather than judicial authorities, become the arbiters of lawful expression. They should avoid delegating responsibility to companies as adjudicators of content, which empowers corporate judgment over human rights values to the detriment of users.

69. States should publish detailed transparency reports on all content-related requests issued to intermediaries and involve genuine public input in all regulatory considerations.

Recommendations for ICT companies

70. Companies should recognize that the authoritative global standard for ensuring freedom of expression on their platforms is human rights law, not the varying laws of States or their own private interests, and they should re-evaluate their content standards accordingly. Human rights law gives companies the tools to articulate and develop policies and processes that respect democratic norms and counter authoritarian demands. This approach begins with rules rooted in rights, continues with rigorous human rights impact assessments for product and policy development, and moves through operations with ongoing assessment, reassessment and meaningful public and civil society consultation. The Guiding Principles on Business and Human Rights, along with industry-specific guidelines developed by civil society, intergovernmental bodies, the Global Network Initiative and others, provide baseline approaches that all Internet companies should adopt.

71. The companies must embark on radically different approaches to transparency at all stages of their operations, from rule-making to implementation and development of "case law" framing the interpretation of private rules. Transparency requires greater engagement with digital rights organizations and other relevant sectors of civil society and avoiding secretive arrangements with States on content standards and implementation.

72. Given their impact on the public sphere, companies must open themselves up to public accountability. Effective and rights-respecting press councils worldwide provide a model for imposing minimum levels of consistency, transparency and accountability to commercial content moderation. Third-party non-governmental approaches, if rooted in human rights standards, could provide mechanisms for appeal and remedy without imposing prohibitively high costs that deter smaller entities or new market entrants. All segments of the ICT sector that moderate content or act as gatekeepers should make the development of industry-wide accountability mechanisms (such as a social media council) a top priority.